Quan Nguyen

ASCII:

- ASCII abbreviated from American Standard Code for Information Interchange, is a character encoding standard for electronic communication [4]

- Overview: Originally based on the English alphabet, ASCII encodes 128 specified characters into seven-bit integers. Ninety-five of the encoded characters are printable: these include the digits 0 to 9, lowercase letters a to z, uppercase letters A to Z, and punctuation symbols [4]

- Old character encodings such as ASCII are from the (pre-) 8-bit era, and try to cram the dominant language in computing at the time, i.e. English, into numbers ranging from 0 to 127 (7 bits). With 26 letters in the alphabet, both in capital and non-capital form, numbers and punctuation signs, that worked pretty well. ASCII got extended by an 8th bit for other, non-English languages, but the additional 128 numbers/code points made available by this expansion would be mapped to different characters depending on the language being displayed. [3]

- Variants and derivations: UTF-8, Unicode [4]

Unicode:

- Unicode and the ISO/IEC 10646 Universal Character Set (UCS) have a much wider array of characters and their various encoding forms have begun to supplant ISO/IEC 8859 and ASCII rapidly in many environments. While ASCII is limited to 128 characters, Unicode and the UCS support more characters by separating the concepts of unique identification (using natural numbers called code points) and encoding (to 8-, 16-, or 32-bit binary formats, called UTF-8, UTF-16, and UTF-32, respectively). [4]

UTF-8:

- UTF-8 is one encoding of Unicode with all its codepoints [1]

- In UTF-8, a character can consist of more than one byte [1]

- UTF-8 is one of the Unicode Transformation Formats which convert a Unicode “codepoint” or hexadecimal integer into a particular sequence of bytes. UTF-8 encodes Unicode characters into a sequence of 8-bit bytes [2]

- UTF-8 is a character encoding - a way of converting from sequences of bytes to sequences of characters and vice versa [3]

- UTF-8 is a particular way of encoding Unicode [3]

Latin1:

- It is an encoding

- Latin1 encodes less than 256 characters [1]

- In latin1 each character is exactly one byte long [1]

- Eight-bit extensions of ASCII, (such as the commonly used Windows-ANSI codepage 1252 or ISO 8859-1 “Latin -1”) contain a maximum of 256 characters.2 [2]

- The ISO-8859 standards are the most common forms of this mapping; ISO-8859-1 and ISO-8859-15 (also known as ISO-Latin-1, latin1, and yes there are two different versions of the 8859 ISO standard as well) [3]

- It is part of the ISO/IEC 8859 series of ASCII-based standard character encodings [5]

- It is the basis for some popular 8-bit character sets and the first two blocks of characters in Unicode [5]

How are emoji's encoded?

- In the Unicode Standard, each emoji is represented as a "code point" (a hexadecimal number) that looks like U+1F063, for example. Thanks to Unicode, our devices all over the world can all agree that U+1F603 is the combination that triggers a grinning face. [6]

Source:

[1]: https://stackoverflow.com/questions/2708958/differences-between-utf8-and-latin1

[2]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480066/

[3]: https://stackoverflow.com/questions/643694/what-is-the-difference-between-utf-8-and-unicode

[4]: https://en.wikipedia.org/wiki/ASCII

[5]: https://en.wikipedia.org/wiki/ISO/IEC\_8859-1

[6]: https://www.codenewbie.org/blogs/what-emojis-tell-us-about-encoding